HYPER-MULTI-STEP: THE TRUTH BEHIND DIFFICULT LONG-CONTEXT TASKS

Yijiong Yu[a]

[a]Tsinghua University, yuyj22@mails.tsinghua.edu.cn

ABSTRACT

Long-context language models (LCLM), characterized by their extensive context window, is becoming increasingly popular. Meanwhile, many long-context benchmarks present challenging tasks that even the most advanced LCLMs struggle to complete. However, the underlying sources of various challenging long-context tasks have seldom been studied. To bridge this gap, we conduct experiments to indicate their difficulty stems primarily from two basic issues: "multi-matching retrieval," which requires the simultaneous retrieval of multiple items, and "logic-based retrieval," which necessitates logical judgment within retrieval criteria. These two problems, while seemingly straightforward, actually exceed the capabilities of LCLMs because they are proven to be hyper-multi-step (demanding numerous steps to solve) in nature. This finding could explain why LLMs struggle with more advanced long-context tasks, providing a more accurate perspective for rethinking solutions for them. Our code and datasets are publicly available at https://github.com/yuyijiong/hard_retrieval_for_llm.

1 INTRODUCTION

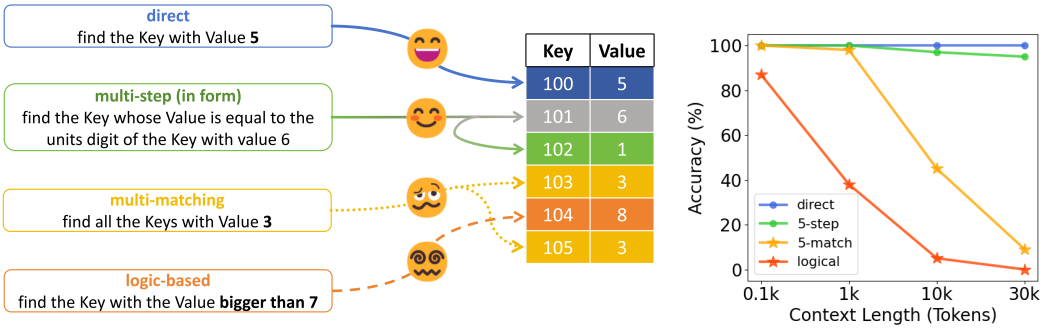
In the past year, long-context language models (LCLMs) such as GPT-4o-128k (OpenAI, 2023) and Gemini-1.5-1000k (Team et al., 2023) have surged in popularity, raising questions about their efficacy in handling extended context tasks. While various LCLMs have demonstrated excellent long-context retrieval ability by passing the "Needle in a Haystack" test (gkamradt, 2023) in over 100k context length, benchmarks like Loogle (Li et al., 2023) and Loong (Wang et al., 2024b) have highlighted their shortcomings in more complex tasks.

To better emulate real-world challenges, recent long-context benchmarks predominantly aim to enhance the difficulty of long-context tasks by requiring real-life scenarios understanding and multi-step processes or increasing the volume of information that needs to be aggregated. This lead these tasks to usually engage multiple capabilities of LLMs, such as world knowledge, advanced reasoning skills and retrieval capability. Consequently, poor model performances are often vaguely attributed to inadequate understanding, reasoning and retrieval capabilities in long context, making it challenging to identify the specific factors that constitute the primary difficulty.

To reveal the real source of challenges in long context tasks, we conduct a detailed analysis (detailed in Appendix C.2 and C.3) of challenging tasks from previous long-context benchmarks, and as expected, we identify 2 common factors that make them difficult: multi-matching retrieval and logic-based retrieval. Multi-matching retrieval involves recalling multiple items simultaneously, and logic-based retrieval involves logical judgment within retrieval criteria. Although they are both "basic" retrieval problems having a straightforward form and cannot be explicitly decomposed into multiple steps using Chain-of-Thought (Wei et al., 2022) (in Appendix C.1, we use examples to detail their differences from those traditional multi-step tasks which can be decomposed by CoT, hence called "formally multi-step"), our experiments, as shown in Figure 1, demonstrate they are much harder for current LCLMs as the context length grows, compared to direct retrieval or formally multi-step retrieval.

Rather than merely focusing on the superficial phenomena, we endeavor to elucidate why these ostensibly simple issues present substantial challenges for LLMs. Through more in-depth experiments, we demonstrate that they are "hyper-multi-step" in nature, which are quite distinct from

normal retrieval problems. "Hyper-multi-step", the truth behind difficult long-context tasks, refers to a problem that appears indivisible in form but actually requires numerous independent steps, and the number of steps will increase indefinitely with the length of the context, that exceed the capacity of LLMs to process simultaneously. To date, none of the techniques such as Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT) prompting and LCLMs have adequately addressed such problems.



(a) different retrieval problems

(b) accuracy of retrieval problems

Figure 1: (a) Examples of direct, formally multi-step, multi-matching and logic-based Key-Value retrieval. (b) Accuracy of GPT-4o in different KV retrieval tasks, as the context length increases. 5-step means a multi-step retrieval task requiring at least 5 times of retrieval. 5-match means a multi-matching retrieval task with 5 matching items for one query. Logical means logic-base retrieval.

Through our studies, we do not aim to propose more challenging benchmarks to drive further optimizations in LCLMs. Rather, we demonstrate a tough reality: while LCLMs are intended to process vast amounts of input data simultaneously, there exist certain long-context tasks which always remain unattainable for LCLMs to solve within one step. Therefore, future research should focus on addressing the challenges associated with numerous steps, rather than merely extending the context window of LLMs.

Our major contributions are as follows:

- We summarize previous long-context benchmarks and evaluate current LCLMs to identify 2 common factors that greatly contribute to the difficulty of long-context tasks: multi-matching retrieval and logic-based retrieval.
- We propose that the essence of the 2 problems is hyper-multi-step, and we provide detailed proof and explanations for this assertion. This was never specified in previous researches.
- We prove that LCLMs inherently have limitations, offering new insights for understanding and addressing long-context problems.

2 Related Works

2.1 Long-Context Language Models

The emergence of long-context language models (LCLMs) aims to enable language models to handle vast amounts of input information simultaneously. In recent years, closed-source LLMs have pioneered advancements in long-context modeling, with context windows expanding from 128k to 1000k tokens. Notable models include GPT-4o (OpenAI, 2023), Claude3.5-200k (Anthropic, 2024), and Gemini-1.5-1000k (Team et al., 2023), which are capable of processing significantly longer texts. Concurrently, open-source models such as phi-3.5-mini (Abdin et al., 2024) and Qwen2.5 (Team, 2024) leverage advanced RoPE (Su et al., 2021) interpolation techniques like Yarn (Peng et al., 2023) and LongRope (Ding et al., 2024) to achieve a 128k context window. These open-source models are usually extended from a 4k pretraining length through long-context post-training with interpolated RoPE. However, it remains to be seen whether these models can truly achieve accurate and efficient handling of lengthy contexts.